Comparing Live and VideoObservation to Assess Early Parent-child Interactions in the Home

Gridley, Nicole; Bywater, Tracey-Jane; Hutchings, Judith

Journal of Child and Family Studies

DOI:
10.1007%2Fs10826-018-1039-y

Published: 01/06/2018

Peer reviewed version

Title Page w/ ALL Author Contact Info.

Comparing live and video observation to assess early parent-child interactions in the home

Nicole Gridley and Tracey Jane Bywater

Department of Health Sciences, University of York, York, UK, YO10 5DD

Judy Mary Hutchings

Centre for Evidence Based Early Intervention, Bangor University, Bangor, UK

This study was conducted as part of a self-funded Masters and a fully funded PhD (2/3 School of Psychology, Bangor University; 1/3 Children's Early Intervention Trust Charity). We would like to acknowledge Dr's Karen Jones, Nia Griffith, Catrin Eames, Pamela Martin-Forbes, Kirsty Pye and Joanna Charles for conducting the live observation visits. Finally, we would like to thank the parents and children who took part in the randomised controlled trial from which this data was drawn, for their commitment to the project and for providing us with this data.

Correspondence concerning this article should be addressed to Nicole Gridley, Department of Health Sciences, University of York, York, UK, YO10 5DD.

Email: nicole.gridley@york.ac.uk

Abstract = 199/250 words

Main Text word count (exc. figures/tables): 5168 without references

Abstract

Observation is the 'gold standard' for assessing parent-child behavior, however few studies have compared coding live, in real time, versus coding from videotapes in terms of their achievable levels of coder reliability within the field of parent programme research. This is important for practitioners and researchers for whom decisions might be influenced by time and financial constraints, but where outcomes may have real practical and clinical implications. Trained coders in the Dyadic Parent-Child Interaction Coding System Revised, coded 40 half-hour videotapes of 33 parent-toddler dyads interacting in the home on 29 items of dyadic behaviour. Four theorised composite variables were constructed. Videotaped data were compared to data drawn from the same interactions previously coded 'live' in the home. Correlations indicated significant agreement between the two modes at the item by item level ($p < .001$). Wilcoxon Rank tests revealed significant differences ($p < .001$) between the two modes. Eight items exceeded a ±30% change in median score suggesting clinically relevant differences. Although both methods achieved acceptable levels of inter-rater reliability, video coding achieved higher levels of agreement. Subtle differences exist between the two modes. Whilst neither mode proved superior it is suggested that they should not be used interchangeably.

Key words: Observation, Reliability, Agreement, Parent-Child Interaction

In the absence of genetic or organic disorders disruption in the very early bond between mother/parent and child is key to understanding many of the social, emotional and/or behavioural problems that arise in later childhood (Hart & Risley, 1995; Leadsom, Field, Burstow & Lucas, 2013; Melhuish et al., 2012). Observations of early parent-child interactions can identify initial problems within the dyadic relationship, detecting families that may benefit from interventions such as parenting support, before problematic child behaviours become engrained (Gardner, 2000; Hawes & Dadds, 2006). In traditional observational methodology, the observer has always been the key 'research instrument', coding behaviour as it happens '*live'* in real-time (Gardner, 2000). However, the increasing availability of digital technology, at a reasonable cost, has provided the opportunity for practitioners and researchers to utilise *video* recordings. Few studies have directly compared modes of observation (i.e. *video* versus traditional *live*) and those that have present conflicting findings (Curby, Johnson, Mashburn & Carlis, 2016; Elsen, Hersen & Agras, 1973; Fagot & Hagan, 1988; Kent, O'Leary, Dietz & Diament, 1979; Moore & Lee, 1974). Comparing live and video modes of observation has not yet been addressed within family or early years research, despite implications for treatment and interventions offered.

The ability of the parent to respond sensitively to their child's needs at an age appropriate level, promotes the development of good child social and emotional competency (Fernald, Marchman & Weisleder, 2014; Huttenlocher, Waterfall, Vasilyeva, Vevea & Hedges, 2010), and language skills (Gridley, Baker-Henningham, & Hutchings, 2016; Hart & Risley, 1995). In contrast, persistent harsh, inconsistent and unresponsive parenting, creates confusion for young children as they are not able to learn the effect of their behaviour on those around them (Scott et al. 2014). In chronic cases these children are likely to develop problematic behaviours (Scott et al., 2014), resulting in an additional cost of £70,000 per individual in terms of service use by the time a child reaches their late twenties (Scott, Knapp, Henderson & Maugham, 2001). Early identification of children at risk for developing behaviour problems, due to problematic parenting, enables action to address these problems and facilitate the positive development of the child and family, and benefit the wider community.

Observation is considered the 'gold standard' method for acquiring in-depth information about the dyadic relationship, and identifying any problems that may require intervention (Gardner, 2000; Hawes & Dadds, 2006). In

contrast to other research methods, such as parent self-report or interviews, observational methods are objective and independent from participant response bias and enhance the opportunity to obtain ecologically valid data that reflects actual behaviour (Gardner, 2000; Wysocki, 2015). Consequently, data drawn from observational methodology, particularly when applied in naturalistic settings, helps to identify problems and facilitate the route to intervention (Bennetts, Mensah, Westrupp, Hackworth & Reilly, 2016).

Currently researchers and practitioners have two options for coding and assessing parent-child interactions; *live* in vivo, or from pre-recorded *videos.* Whilst videotaped technology has been available for 40 years, modern technological advancements and relatively low costs of digital equipment have led to researchers and clinicians increasingly using videotape as the preferred means of collecting and analysing observational data (Shrum, Duque & Brown, 2005). Video is considered advantageous to live coding as it permits researchers to replay segments of the interaction to gain a deeper understanding of a series of events whilst maintaining objectivity (Haidet, Tate, Divirgilio-Thomas, Kolanowski & Happ, 2009). In addition, video allows data to be re-used for subsequent research and coded using different frameworks (Ryan et al., 1995). However good quality recordings are reliant on the skill of the observer to accurately position the camera in order to avoid the omission of important behaviours (Gardner, 2000; Haidet et al. 2009; Rosenstein, 2002). In addition, video equipment can be intrusive. This is a particular concern when the environment is cramped, e.g. in some home observations, and where there is also a need for the researcher to remain present during the interaction (Hutchings et al. 2007; Rosenstein, 2002). Finally, videotaped observation is more time consuming and more resource intensive compared to live observations. It can also be subject to mechanical failure, power issues, poor lighting and sound quality (Johnson & Bolstad, 1975). Consequently, when deciding how to collect observational data, e.g. using live or video coding, there is a need to consider the impact that the mode of data collection may have on the data and overall conclusions.

Testing the hypothesis that *live* and *video* modes of observation may produce different outcomes has not yet been studied in family research, although it has previously been the subject of enquiry across several different fields, i.e. educational and social psychology (Curby et al., 2016; Elsen et al., 1973; Fagot & Hagan, 1988; Kent et al., 1979; Moore & Lee, 1974). The primary objective of previous research has been to explore the achievable level of

agreement between coders (inter-rater reliability) rather than to assess the relationship between them. The sparse existing literature displays conflicting results. For example, two studies of school-based behaviours (Fagot & Hagan, 1988; Kent et al., 1979) found greater inter-rater reliability achieved for live observation, whilst two studies of social psychology (Elsen et al., 1973; Moore & Lee, 1974) concluded that live and video modes of observation were not too dissimilar in terms of levels of inter-rater reliability. More recently, Curby et al., (2016) reported that whilst both modes were comparable in terms of achievable levels of inter-rater reliability, video observation produced fewer recorded frequencies of both verbal and non-verbal behaviours. Collectively these results indicate that differences between modes do exist in terms of both levels of reliability and content.

The purpose of the current study was to compare the utility of live and video modes of observation when coding home-based parent-child interactions using a standardised measure of parent-child behaviour. The Dyadic Parent-Child Interaction Coding System – Revised (DPICS-R; Webster-Stratton, 2000) has been used in many evaluation studies of parenting programmes (Hutchings et al., 2007; Hutchings, Griffith, Bywater & Williams, 2017; McGilloway et al., 2014; Reid, Webster-Stratton & Beauchaine, 2001; Seabra-Santos et al., 2016; Webster-Stratton & Hammond, 1997). It is also used routinely by Parent Child Interaction Therapy International. The specific questions to be addressed in the present study were:

1. To what degree are live and video ratings of the same parent-child interaction related to one another?
2. Are there mean differences in live and video ratings of the same interaction?
3. Are there clinically relevant differences in live and video ratings of the same interaction?
4. Is the achievable level of inter-rater reliability similar across both live and video modes of observation?
5. Are there differences in between modes at the composite variable level?

Method

Participants

A total of 89 families participating in a randomised controlled trial (RCT) evaluation of the Incredible Years (IY; Webster-Stratton, 2010) Toddler Programme (see Hutchings et al. 2017 for recruitment procedures and trial outcomes) provided written informed consent to be observed (live) and simultaneously videotaped for 30-minutes during naturalistic free play with their toddlers in their home. Over the course of three research visits conducted six months apart over a 12-month period 192 observations were conducted (Baseline *n* = 89, Follow Up 1 *n* = 67, and Follow Up 2 *n* = 36). For the purposes of the current study a sub-sample of 40 observations were randomly selected by hand by the lead author and included in the analysis. Videotaped observations were selected from the upright DVD case in which they were stored. These observations were organised numerically (by ascending study ID number), and then by time-point (Baseline, Follow Up 1 and Follow Up 2). With the exception of the top DVD, no identifiable information (ID number or time point) was available to the researcher during the selection process.

The final set of observations selected for inclusion in the current study represented 31 of the 89 original families recruited for the RCT. Sixteen cases related to seven families where observations had been conducted at two or more of the three available time points. The remaining observations related to 24 independent families taken at various time points across the RCT. The final dataset relates to a sample of children with a mean age of 27.32 months (*SD* = 9.44), and mothers with a mean age of 29.75 years (*SD* = 6.58). The mean number of people present in the room at each observation, excluding the researcher, was 2.60 (*SD* = 0.84). The primary coder for the current study was present at 55% of all live observations conducted during the RCT. The secondary coder was present for 15% of live observations.

Procedure

As part of the RCT (Hutchings et al., 2017) parents were asked to consent to 1) being observed live in the home, and, 2) to being videotaped. Observers were trained to ≥70% inter-rater reliability using the Dyadic Parent-child Interaction Coding System- Revised (DPICS-R; Webster-Stratton, 2000) before undertaking home visits. During the home visits the parent (the mother in all cases) was observed interacting with their toddler for 30-minutes. There were no specific instructions other than asking the parent to play, as they would normally, with their child and to ensure that the television was off. During this half-hour period one of six experienced DPICS-R users coded the

interaction continuously in six continuous five-minute segments. A camera was placed close to the researcher and simultaneously recorded the interaction for later analysis. Inter-rater reliability visits were conducted across 20% of all live visits.

On completion of the RCT two experienced and reliable (trained to 70% agreement) DPICS-R coders randomly selected by hand 40 videotapes from the larger battery of 192 available videotapes and coded each video using the DPICS-R. To ensure consistency between the two modes coders in the video condition only viewed the videotaped observations once in real time. Videos were viewed in a private room where the two observers independently coded each interaction by recording each time a behavior occurred by marking a tally on a score sheet next to the relevant behaviour item. The continuous recording of all behaviours across the 30-minutes provided frequency counts for each of the 29 DPICS-R items. Inter-rater reliability percentage agreements were established at the end of each 30-minute video. For both modes of observation, live or video, the scores provided by the primary observer were taken as the most accurate and used for the final analysis. The average period between the coding of the live observation and coding the video version was 14.08 (*SD* = 5.35) months hence familiarity with previous coding was unlikely to bias scoring of the videos.

Measures

Dyadic parent-child interaction coding system – revised (DPICS-R, Webster-Stratton, 2000).

The DPICS (Eyberg & Robinson, 1983; Robinson & Eyberg, 1981; DPICS II, Eyberg & Robinson, 2005) is an observational tool designed to assess the quality of parent-child social interaction. Standardised and validated across a variety of settings with varying populations (Bjorseth, McNeil & Wichstrom, 2015; Eyberg & Robinson, 2005; Robinson & Eyberg, 1981) it provides a comprehensive account of behaviour due to its use of continuous recording of interactions using frequency counts. A revised version of the DPICS (the DPICS-R) was devised in 2000 (Webster-Stratton, 2000) for use as the main outcome measure for assessing behavioural change following the implementation of parenting interventions with parents of children aged from 12 months to 12 years (e.g. Hutchings et al., 2007). The DPICS-R consists of 29 parent and child categories (Table 1) each coded as continuous

frequency counts across six five-minute segments. Previous users of the DPICS and the DPICS-R have reported achievable levels of inter-rater reliability for both parent and child individual categories with percentage agreements >70% and ICC's > .67 (Robinson & Eyberg, 1981; Hutchings et al., 2017), and good discriminate validity when used to code live dyadic interactions (Bjorseth, McNeil & Wichstrom, 2015). Whilst there are no formal cut-offs for the DPICS or the DPICS-R the original DPICS developers, Robinson and Eyberg (1981) suggested that a 30% change in scores from the first to the second assessment represented a clinically relevant change.

Four composite variables; Positive Parent, Negative Parent, Child Positives and Child Negatives (see Table 1 for items included under these composite variables) can be derived using 14 of the 29 DPICS-R items and have previously been used by the developer and in evaluation studies of parenting programmes as important indicators of change (Bywater et al., 2009; Jones, Daley, Hutchings, Bywater & Eames, 2007; Robinson & Eyberg, 1981). These variables have not, however, been subject to robust statistical testing.

(Table 1 here)

Data Analysis

The continuous frequency counts for each of the 29 DPICS-R items and their associated composite variables (in both live and video) were subjected to normality tests in SPSS 23.0. Twenty of the 29 individual items were non-normally distributed in either mode of observation. To enable analysis across all items in both modes of observation non-parametric tests were adopted at the item by item level. Inspection of frequency distributions indicated that Parent Ignores, Grandma's Rules and Warnings (see Table 1 for category descriptions) were low in frequency (≤1) in both modes of observation. As a result, these categories were excluded from further analysis because the analysis would not be meaningful. The four theoretical composite variables (Parent Positive, Parent Negative, Child Positive and Child Negative) demonstrated normally distributed values appropriate for parametric tests.

To establish the degree to which live and video modes of observation were related a series of Spearman's correlations were applied to the remaining 26 individual DPICS-R items due to non-normally distributed items.

Pearson's correlations were applied to the four composite variables for normally distributed items. To establish mean differences in live and video ratings of the same interaction the Wilcoxon Signed Rank test (a non-parametric equivalent of the paired t-test) was applied to each of the 26 individual items using a Bonferroni corrected *p* value of ≤.001 to account for the number of analyses. Paired t-tests were then applied to the four composite variables.

To establish whether the mode of observation changed outcomes at a level considered to be clinically relevant, exploratory analysis was undertaken using a pre-defined ±30% threshold change in median score as the criterion for assessment. As video technology is the newer method for conducting observation, live observation codes were used as the gold standard against which to compare video scores. The upper and lower 30% thresholds of the live median scores were calculated and the video median scores were then compared against this range. Decision rules for determining clinically relevant differences in scores associated with mode of observation were as follows; median video scores that fell within the ±30% range indicated that no clinically relevant difference in scores had occurred. Median video scores that exceeded this threshold in either direction suggested a clinically relevant change in score had occurred.

Inter-rater reliability between coders in both modes was assessed at two levels; using percentage agreements at the global level and Intra-Class Correlations (ICC) using a two-way mixed model with absolute agreement at the item by item and composite variable level. Finally, to establish the internal consistency of the four theorized composite variables a series of Cronbach's alpha were applied.

To ensure that the inclusion of multiple observations from the same family did not impact upon the conclusions drawn the main analysis (correlations, t-tests and clinically relevant differences) was conducted twice; firstly, using the full sample of 40 observations, and then using only one observation point for each of the 31 independent families. In the 16 cases where there was more than one video per family, only the earliest observation was included in the analysis.

Results

Research Question 1: Correlations Between Modes of Observation

Using the full sample of 40 observations Spearman's correlations at the item by item level (Table 2) indicated that 17 of the 19 parent categories and all of the seven child categories coded using the live mode of observation were statistically related ($p \leq .001$ Bonferroni correction) to the same category when coded using video observation. These positive correlations ranged from moderate to large ($r = .565$ to .962). Two parent items coded live (physical intrusions and physical negatives) were not statistically related to the video codes of the same categories.

Using the smaller sample of 31 observations only one category 'child positive affect verbal' gave different results. Video coded child positive affect verbal was no longer statistically related to the its live counterpart ($r = .342$, $p = .060$). All other results remained the same.

(Table 2 here)

Pearson's correlations were applied at the composite variable level for normally distributed data (Table 3). Results using both the full ($N = 40$), and the smaller sample of 31 cases indicated that all four theorised composite variables (positive parent, negative parent, child positive and child negative) were positively related at a moderate level. These findings suggest that a relationship does exist between the live and video scores of the same individual and composite variable categories of the DPICS-R.

(Table 3 here)

Research Question 2: Mean Differences Between Modes of Observation

A series of Wilcoxon Signed Rank tests were conducted at the item by item level to assess differences between live and video mean rank ratings of the same parent and child interaction. Results using the full sample ($N = 40$) indicated that seven parent items (physical intrusions, physical negatives, physical positives, descriptive questions, descriptive comments, indirect commands and direct commands) and four child items (physical negatives, cry/whine/yell, positive affect non-verbal and physical warmth) were statistically different when coded using the two different modes of observation (Table 2). However, once the Bonferroni correction had been applied only two

parent (descriptive comments and direct commands) and two child items (cry/whine/yell and positive affect non-verbal) remained statistically different.

Analysis using the smaller sample (n = 31) altered the findings for three parent categories. Physical intrusions ($t(30)$= -3.398, p = .001) and indirect commands ($t(30)$= -3.252, p = .001) indicated statistically significant differences between modes, whilst the category of direct commands was no longer statistically different ($t(30)$= -2.745, p = .006)

To test for differences between modes at the composite variable level a series of paired t-tests were applied (Table 3). Results using both the full and smaller sample indicated that child positives was the only composite variable to indicate a statistical difference in scores between the two modes. These results suggest that overall there is very little difference in ratings of parent and child interactions when using either live or video modes of observation.

Research Question 3: Clinically Relevant Difference Between Modes

Using the live observation median scores as the gold standard of observation a predefined ±30% change in score was used as a preliminary guide to assess the impact of observational mode on outcome at a clinically relevant level. Video codes for four of the 19 parent and three of the seven child categories indicated scores that exceeded the ±30% criteria applied to the live scores. Parent physical intrusions and child cry/whine/yell's coded using the video mode demonstrated scores that exceeded the upper 30% threshold of the live scores i.e. more of these behaviours were recorded using video. Whilst video codes for parent physical positives, verbal questions, reflective questions, child smart talk and child positive affect non-verbal demonstrated scores that exceeded the lower 30% threshold of the live codes scores i.e. less of these behaviours were recorded using video.

Using the smaller sample of 31 videos three changes from the larger sample analysis were observed. Firstly, the frequency of parent negative commands coded using video were shown to exceed the clinical threshold in a negative direction i.e. less were observed. Secondly, codes for parent physical positives in the video condition were no longer clinically different from those observed in the live condition in terms of ±30% above the median. Finally,

video coded child smart talk was no longer clinically different to its live counterpart. All other clinically relevant differences were upheld suggesting some subtle differences between the two modes in capturing parent-child interactions in the home with an impact for some categories at a clinically relevant level.

Research Question 4: Inter-Rater Reliability within Modes of Observation

To allow for comparability across modes, inter-rater reliability checks were conducted on those videos where the same observation had been a live inter-rater reliability visit. From the 40 randomly selected videos coded for this study 30% (*N* = 12) had been a live inter-rater reliability visit.

An assessment of the overall percentage agreement between coders within the live mode of observation indicated an overall mean agreement of 73.67% (*SD* = 17.80). Agreement between coders ranged between 43% and 98%. In contrast, coders using the video mode of observation indicated an overall agreement of 89.75% (SD = 8.45), with a range of 67-99%. These results suggest desirable levels (≥70%) of inter-observer reliabilities can be attained using percentage agreements in either mode of coding. However, greater levels of inter-observer reliability were consistently attained when employing the video mode of coding.

A comparison of inter-observer reliability was also conducted at the item by item and theorised composite variable level using ICC's with a two-way mixed model with absolute agreement. Results from the item by item level analysis using the live mode of observation indicated that 15 parent and five child categories and all four composite variables were significantly correlated. However, once the Bonferroni correction was applied only nine parent (Physical Intrusion, Physical Positive, Unlabeled Praise, Acknowledgment, Verbal Question, Reflective Statement, Statement, Indirect and Direct Command), one child category (Cry/Whine/Yell), and three composite variables (Positive Parent, Negative Parent and Child Negatives) were significantly correlated. In comparison, data captured using the video mode of observation indicated that 19 parent and five child categories and all four composite variables yielded statistically significant results.

These findings indicate that a greater level of inter-rater reliability can be achieved at both the item by item level, and at the theorised composite variable level when using video modes of observation in comparison to live.

(Table 4 here)

Research Question 5: Internal Consistency of Theorised Composite Variables

The four composite variables of the DPICS-R are theorised constructs that have been applied in research as a way of reducing the data to something more theoretically meaningful for analysis, and have been shown to demonstrate meaningful post-intervention change. Despite this, the authors are not aware of any previous statistical analysis being conducted to assess these constructs for their statistical robustness. Although the assessment of the factor structure of DPICS-R is beyond the scope of this study a series of Cronbach Alpha's were applied to the four theorised composite variable to establish the level of internal consistency using live and video observational ratings. Results (see Table 5) indicated that irrespective of observational mode the internal consistency for each of the four composite variables was unable to reach an acceptable level (live α range = .236 to .601; video α range = .059 to .436). These findings suggest that the individual items that are used to form these four theoretical composite variables may measure different underlying constructs and that further investigation of the DPICS-R items, and how these may be reduced to more meaningful components for analysis i.e. using exploratory factor analysis, is warranted.

(Table 5 here)

Discussion

This study compared live and video observational modes of coding parent-child interactions to assess their agreement when applying a complex coding system (DPICS-R) used routinely in research and clinical practice. The latest version of the DPICS coding scheme (DPICS-III) is one of the preferred measures used by Parent Child Interaction Therapy International, a high profile organisation whose purpose is to promote fidelity and evidence based practice within the field of family functioning. Results from the current study using the DPICS-R demonstrated high levels of agreement between the two modes at both the item by item level and composite variable level. Several items using video coding were shown to exceed a proposed ±30% threshold from the live median scores suggesting some potentially clinically relevant differences between the two modes at the item level.

In addition, although good levels of inter-rater reliability were achieved in both modes, coder agreement using video was generally higher and more consistent. The findings suggest that there is agreement between the two modes of observation in both outcome and achievable levels of reliability, however caution should be taken if considering using the two modes interchangeably.

The current study did not find sufficient evidence to suggest that the scores from the two modes were significantly different from one another. However, what may constitute a clinically relevant difference may not be reflected in a statistically significant difference in research (Middel & van Sonderen, 2002; Page, 2014). For this reason, and in line with previous recommendations from the original developers of the DPICS (Eyberg & Robinson, 1981) we imposed a ±30% threshold using the live median scores as the gold standard to establish whether the video scores indicated a difference in proportion of cases of potential clinical relevance. The findings indicated that only a handful of items when coded using video exceeded this ±30% threshold and these items were observed to be those that occurred in low frequency. The DPICS-R does not have specified clinical cut-offs and the imposed ±30% threshold may not actually constitute a clinically relevant difference in practice for the items occurring in low frequency. Further validation work, with other gold standard measures of parent and child behavior, is required to establish clinical cut-offs for the DPICS-R, and thus determine the relevance of level of agreement between the two modes. The results could be useful to determine thresholds for referring families to specific interventions.

Limitations

One weakness of this study is the sampling procedure. Due to time constraints, it was not possible to use a rigorous method of random selection and hand selection was imposed. Notwithstanding the lack of rigor associated with this selection procedure it was considered that this technique was appropriate given that each observation stood an equal chance of being included in the current study. Despite this, because videos were selected from observations recorded at all three time points in the study, the final sample of 40 observations included multiple interactions drawn from the same families (*n* = 16 videos) introducing noise into the data. To control for this bias the analysis was re-run to include only one observation from each of the 31 families. The re-

analysis indicated relatively few changes to the overall findings confirming that, in spite of the sampling method employed, the two modes do not differ significantly from one another.

A second limitation is that data drawn from the live observational visits were coded independently during a RCT, by a pool of six trained observers compared to only two observers in the video condition. Although coders were trained to ≥70% reliability and engaged in fortnightly coding meetings to ensure reliability levels were maintained during busy coding periods, variability across coders or amongst coding pairs may be a possible confounding variable within our data. This may explain the variability within the live inter-rater agreement and the consequent differences between the two modes on inter-rater reliability. It is suggested that future coders should achieve a higher level of reliability before coding commences, and that frequent supervision or reliability checks are undertaken between coders to ensure standards are maintained and that coder drift is minimised.

Finally, the primary coder for the present study was present at 55% of all live observations conducted for the RCT and the secondary coder present at 15%. Consequently, it is possible that there may have been carry over effects associated with familiarity for the video mode of coding which may have contributed to the positive correlations between the two modes. However, video coding took place approximately 14 months after the original live coding, and many observations were conducted in the intervening period suggesting that this may have been unlikely.

Strengths

The main strength is that this is the first known study of its kind within family research, specifically with parents of children within this pre-school age group. Consequently, it is novel and has the potential to generate further research in this area. Moreover, the DPICS has been used in a number of evaluations of parenting programmes, and the most recent edition (DPICS-III) is routinely used to evaluate PCIT. As a result, the findings from this study have both clinical and practical implications for practitioners and researchers specifically in terms of informing decisions of how to best use time and financial resources effectively. However, further research in this area is required before these recommendations can be applied to practice.

Secondly, the current data were collected as part of a rigorous RCT by skilled researchers, proficient in the use of the DPICS-R, with extensive knowledge of an array of standardised and validated coding systems. The researchers were experienced in carrying out home observations and controlling technical equipment whilst conducting detailed live observations. The 40 videos used for analysis were not subject to any mechanical failures and comprised 'clean' and audible 30-minute interactions, although they were not randomly selected on this basis.

The final strength of this study is that the coders in video mode only viewed the videotapes once, in real time, to ensure consistency and comparability across the two modes in terms of coding processes. Whilst one strength of videotaped observation is that segments of dense interaction can be re-watched and replayed, we imposed this rule to ensure that any differences that materialised were artifacts of the mode under study and not a byproduct of different coding processes. As a result, we are confident that the results reflect real differences between modes.

The findings have implications for future research for both clinicians and researchers. Firstly, although high levels of inter-observer agreement were attained using both modes of observation, video coding yielding greater and more consistent reliability. Although video coding is likely to be a more expensive and time-consuming alternative to live observation, these results demonstrate that there are pay offs for choosing this mode. For example, greater inter-rater reliability and the ability to subsequently recode the data using different coding schemes, can significantly outweigh the disadvantage of added expense, data protection and storage issues. Moreover, given that clinicians and researchers are increasingly using video technology to conduct observations, these results support the continued use of this mode over more traditional techniques. Further work, however, is needed to establish whether subtle differences between modes result in clinically relevant differences.

The finding that the four theorised composite variables yielded poor levels of internal consistency suggests that the items that make up these composites might be measuring different constructs, and that other items of the DPICS-R might be more highly correlated and form more appropriate constructs. As a result, future exploration of the underlying structural validity of the DPICS-R using data reduction techniques, such as exploratory factor analysis, is recommended to establish and confirm its underlying constructs. Such investigations were beyond the

scope of the present study; however, this is an important area of research particularly if the DPICS-R continues to be extensively used and adapted within evaluation research where findings are of clinical interest.

Ethics Statement

The authors confirm compliance with ethical standards

Conflict of Interest

The authors declare they have no conflicts of interest

Author Contributions

NG designed and executed the study, conducted data analyses, and wrote the paper. TB collaborated with the design and writing of the study, and editing the final manuscript. JH collaborated with the design and writing of the study, and editing the final manuscript.

References

Bennetts, S. K., Mensah, F. K., Westrupp, E. M., Hackworth, N. J., & Reilly, S. (2016). The agreement between parent-reported and directly measured child language and parenting behaviours. *Frontiers in Psychology, 7,* 1-18. doi: 10.3389/fpsyg.2016.01710.

Bjorseth, A., McNeil, C., & Wichstrom, L. (2015). Screening for behvaioural disorders with the Dyadic Parent-Child Interaction Coding System: sensitivity, specificity, and core discriminative components. *Child & Family Behaviour Therapy, 37,* 20-37. doi: 10.1080/07317107.2015.1000228

Bywater, T., Hutchings, J., Daley, D., Whitaker, C., Yeo, S. T., Jones, K., Eames, C., & Tudor Edwards, R. (2009). Long-term effectiveness of a parenting intervention for children at risk of developing conduct disorder. *The British Journal of Psychiatry, 195,* 318-324. doi:10.1192/bjp.bp.108.056531

Curby, T. W., Johnson, P., Mashburn, A. J., & Carlis, L. (2016). Live Versus Video Observations: Comparing the Reliability and Validity of Two Methods of Assessing Classroom Quality. *Journal of Psychoeducational Assessment*, *34*(8), 765-781. doi: 10.1177/0734282915627115

Elsen, R. M., Hersen, M., & Agras, W. S. (1973). Videotape: a method for the controlled observation of non-verbal interpersonal behaviour. *Behaviour Therapy, 4*, 420-425. doi:10.1016/S0005-7894(73)80123-6.

Eyberg, S. M., & Robinson, E. A. (1983). *The Dyadic Parent-Child Interaction Coding System Edition 1.* Unpublished manuscript. University of Washington: The Parenting Clinic.

Eyberg, S. M., & Robinson, E. A., (2005). *The Dyadic Parent-Child Interaction Coding System Edition 2*. Unpublished manuscript. University of Washington: The Parenting Clinic.

Fagot, B., & Hagan, R. (1988). Is what we see what we get? Comparisons of taped and live observations. *Behavioural Assessment, 10,* 367-374. doi: 0191-5401/88.

Fernald, A., Marchman, V. A., & Weisleder, A. (2014). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16,* 234–248. doi:10.1111/desc/12019

Gardner, F. (2000). Methodological issues in the direct observation of parent-child interaction: do observational findings reflect the natural behaviour of participants? *Clinical Child & Family Psychology Review, 3,* 185-198. doi: 1096-4037/00/0900-0185.

Gridley, N., Baker-Henningham, H., & Hutchings, J. (2016). Measuring parental language to target families for early intervention services. *Child Care in Practice, online first.* doi: 10.1080/13575279.2016.1188761

Haidet, K. K., Tate, J., Divirgilio-Thomas, D., Kolanowski, A., & Happ, M. B. (2009). Methods to improve reliability of video recorded behavioural data. *Research in Nursing & Health, 32,* 465-474. doi:10.1002/nur.20334.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hawes, D. J., & Dadds, M. R. (2006). Assessing parenting practices through parent-report and direct observation during parent-training. *Journal of Child & Family Studies*, *15*(5), 554-567. doi: 10.1007/s10826-006-9029

Hutchings, J., Griffith, N., Bywater, T., & Williams, M. (2017). Evaluating the Incredible Years Toddler Parenting Programme with parents of toddlers in disadvantaged (Flying Start) areas of Wales. *Child: Care, Health & Development, 43,* 104-113. doi: 10.1111/cch/12415

Hutchings, J., Bywater, T., Daley, D., Gardner, F., Whitaker, C., Jones, K., Eames, C., Edwards, R. T. (2007). Parenting intervention in Sure Start services for children at risk of developing conduct disorder: pragmatic randomised controlled trial. *British Medical Journal, 334*, 678-684. doi: 10.1136/bmj/j.39126.620799.55.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343-365. doi: 10.1016/j.cogpsych.2010.08.002

Johnson, S. M., & Bolstad, O. D. (1975). Reactivity to home observation: a comparison of audio recorded behaviour with observers present or absent. *Journal of Applied Behaviour Analysis, 8*, 181-185. doi:10.1901/jaba.1975.8-181.

Jones, K., Daley, D., Hutchings, J., Bywater, T., & Eames, C. (2007). Efficacy of the Incredible Years Programme as an early intervention for children with conduct problems and ADHD: long-term follow up. *Child: Care, Health & Development, 33*(6), 749-756. doi: 10.1111/j.1365-2214.2008.00817.x

Kent, R. N., O'Leary, D., Dietz, A., & Diament, C. (1979). Comparison of observational recordings in vivo, via mirror, and via television. *Journal of Applied Behaviour Analysis, 12,* 517-522. doi:10.1901/jaba.1979.12-517.

Leadsom, A., Field, F., Burstow, P., & Lucas, C. (2013). *The 1001 critical days: The importance of the conception to age two period*. London: A Cross Party Manifesto.

McGilloway, S., NiMhaille, G., Bywater, T., Leckey, Y., Kelly, P., Furlong, M., Comiskey, C., O'Neill, D., & Donnelly, M. (2014). Reducing child conduct disordered behaviour and improving parent mental health in disadvantaged families: a 12-month follow up and cost analysis of a parenting intervention. *European Child & Adolescent Psychiatry, 23.* 783-794. doi: 10.1007/s00787-013-0499-2

Melhuish, E., Quinn, L., Sylva, K., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2012). Preschool affects longer term literacy and numeracy: results from a general population longitudinal study in Northern Ireland. *School Effectiveness & School Improvement: An International Journal of Research, Policy & Practice, 24.* 234-250. doi: 10.1080/09243453.2012.749796

Middel, B., & van Sonderen, E. (2002). Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *International Journal of Integrated Care, 2.* doi: 10.5334/ijic.65

Moore, L. F., & Lee, A. J. (1974). Comparability of interviewer, group and individual interview ratings. *Journal of Applied Psychology, 59,* 163-167. doi:10.1037/h0036518.

Page, P. (2014). Beyond statistical significance: clinical interpretation of rehabilitation research literature. *International Journal of Sports Physical Therapy, 9.* 726-736.

Reid, M. J., Webster-Stratton, C., & Beauchaine, T. P. (2001). Parent training in head start: a comparison of program response among African American, Asian American, Caucasian, and Hispanic mothers. *Prevention Science, 2.* 209-227. doi: 10.1023/A:1013618309070

Robinson, E. A., & Eyberg, S. M. (1981). The dyadic parent-child interaction coding system: standardization and validation. *Journal of Consulting & Clinical Psychology, 49,* 245-250. doi: 0022-006x/81/4902-0245.

Rosenstein, B. (2002). Video use in social science research and program evaluation. *International Journal of Qualitative Methods, 1*(3), 1-38. Retrieved from http://www.ualberta.ca/~iiqm/backissues/1_3Final/html/rosenstein.html.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80,* 664-670. doi:0021-9010/95.

Scott, S., Knapp, M., Henderson, J., & Maugham, B. (2001). Financial cost of social exclusion: follow up study of antisocial children into adulthood. *British Medical Journal, 323.* 1-5

Scott, S., Doolan, M., Beckett, C., Harry, S., Cartwright, S., & the HCA team. (2014). *How is parenting style related to child anti-social behaviour? Preliminary findings from the Helping Children Achieve study.* Research Report DFE-RR185a. Retrieved September 2015 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/197732/DFE-RR185a.pdf

Seabra-Santos, M. J., Gaspar, M. F., Azevedo, A. F., Homem, T. C., Guerra, J., Martins, V., & Leitao, S. (2016). Incredible Years parent training: what changes, for whom, how, from how long? *Journal of Applied Developmental Psychology, 44.* 93-104. doi: 10.1016/j.appdev.2016.04.004

Shrum, W., Duque, R., & Brown, T. (2005). Digital video as research practice: Methodology for the Millennium. *Journal of Research Practice, 1.*

Webster-Stratton, C. (2000). *Dyadic parent-child interaction coding system - revised.* Unpublished manuscript. University of Washington. Seattle.

Webster-Stratton, C. (2010). *The Incredible Years Parenting Programme Users Manual.* Unpublished manuscript.

Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: a comparison of child and parent training interventions. *Journal of Consulting & Clinical Psychology, 65,* 93-109. doi: 10.1037/0022-006x.65.1.93.

Wysocki, T., (2015). Introduction to the special issue: direct observation in pediatric psychology research. *Journal of Pediatric Psychology, 40,* 1-7. doi: 10.1093/jpepsy/jsu104

Table 1.
Item descriptions of the DPICS-R

DPICS-R Item Name	Description	Example
Parent Items		
Physical Intrusion	An obtrusive, unsolicited act of entering into or taking over the child's activity or an object which the child is occupied with	*Parent holds down a book that the child is trying to take away*
Physical negative	A touch or bodily contact that inflicts pain, restrains the child or forces/pulls the child	*Parent pulls child up by their wrists and child says 'ouch!'*
Parent Ignore	Child deviant behaviour is ignored for five seconds	*The parent remains silent, maintains a neutral expression, avoids or breaks eye contact with the child*
Critical Statement	A verbalization that finds fault with the activities, products or attributes of the child	*You're being silly now*
Negative Command	An order that tells the child not to do something	*Don't put that on there*
Physical Positive	A touch or bodily contact between the parent and child that is neutral or positive	*Parent strokes child's face*
Positive Affect	A non-verbal expression of enjoyment, warmth or enthusiasm directed at the child that must be seen by the child	*C: smiles at mum* *P: smiles back*
Unlabeled Praise	A non-specific verbalization that expresses a favorable judgement on an activity, product or attribute of the child	*Good girl*
Labelled Praise	A specific verbalization that expresses a favorable judgement on an activity, product or attribute of the child	*Your colouring is beautiful*
Acknowledgement	A brief verbal response to the child's verbalization or behaviour that contains no content other than a simple yes or no	*C: Kitty* *P: Yeah*
Question	A comment expressed in a question form	*Do you have the missing Lego piece?*
Descriptive Question	A question that expresses approval appreciation or positive of the child's effort, attributes or products	*You're getting very good at this aren't you?*
Verbal Question	Any attempt by the parent made to elicit a behavioural response from the child to label objects/people/body parts etc.	*Where's mummy's nose?*
Reflective Question	A statement which repeats all or part of the child's preceding verbalization in question form	C: Boo ball P: You want the blue one?
Reflective Statement	A statement which repeats all or part of the child's preceding verbalisation	*C: Boo ball* *P: Yes, the blue ball*
Statement	A sentence or phrase that gives an account of the objects or people, or activity occurring during the observation	*This is a teddy*
Descriptive Comment	A statement or phrase that describes what the child is doing	*You're putting the cow in the barn*
Verbal Labelling	Any attempt by the parent to label objects/people/body parts etc, whilst holding the child's attention	*(Holding up a yellow crayon) A yellow crayon*
Indirect Command	An order, demand or direction for a behavioural response that is implied, non-specific, or stated in question form	*How about opening the door?*

Direct Command	A clearly stated order, demand or direction in a declarative form i.e. tells the child what to do rather than asks them	*Put the doll in the highchair*
Grandma's Rule	A positive or negative command that specifies a positive consequence if the child complies	*When you put the cars away we can have your favorite treat*
Warning	A statement that includes a positive or negative command accompanied by a negative consequence for non-compliance	*Put the toys away otherwise you will not get a treat*
Child Items		
Physical Negative	A bodily attack or attempt to attack another person	*Child pinches parent/sibling*
Destructive	When the child destroys, damages or attempts to damage any object including animals	*Child throws a Lego block across the room*
Smart Talk	Cheeky or rude speech	*I hate you*
Cry/Whine/Yell	A cry, whine or yell that is deemed as general deviance	*Child screams above room level noise*
Positive Affect Non-Verbal	A non-verbal expression of enjoyment, warmth, or enthusiasm directed at the parent which the parent sees	*Parent smiles* *Child smiles*
Positive Affect Verbal	Positive evaluative verbal expression of pleasure, warmth, enthusiasm or gratitude	*I love you daddy*
Physical Warmth	An explicit physical act of endearment initiated by the child	*Child cuddles into parent*
Composite Variables		
Parent Positives	Summation of all parent Physical Positives, Positive Affect, Unlabelled Praise and Labelled Praise	
Parent Negatives	Summation of all parent Physical Intrusions, Physical Negatives, Critical Statements and Negative Commands	
Child Positives	Summation of all child Positive Affect Non-Verbal, Positive Affect Verbal, and Physical Warmth	
Child Negatives	Summation of all child Physical Negatives, Destructives, Smart Talk and Cry/Whine/Yell's	

Table 2.
Median, range, spearman's correlations and Wilcoxon Signed Rank tests of the individual items of the DPICS-R when coded using live and video modes of observation (*N* = 40)

	Live Median (Range)	**Video Median (Range)**	**± 30%**	***r***	***z***
Parent Items					
Physical Intrusion	1.00 (0-19)	4.00 (0-23)	0.70 to 1.30^	.231	-2.447
Physical Negative	0.00 (0-19)	0.00 (0-6)	-	.315	-2.331
Critical Statement	9.50 (0-42)	10.00 (0-43)	6.65 to 12.35	.843*	-.804
Negative Command	2.50 (0-15)	2.00 (0-16)	1.75 to 3.25	.864*	-.353
Physical Positive	13.00 (1-61)	9.00 (0-55)	9.10 to 16.90^	.866*	-2.716
Positive Affect	22.00 (3-64)	20.50 (4-73)	15.40 to 28.60	.874*	-1.314
Unlabeled Praise	16.00 (0-58)	15.50 (0-55)	11.20 to 20.80	.962*	-1.099
Labelled Praise	0.00 (0-8)	0.00 (0-8)	-	.733*	-.135
Acknowledgement	26.00 (3-113)	27.50 (3-86)	18.20 to 33.80	.674*	-1.238
Question	62.00 (9-151)	76.00 (4-190)	43.40 to 80.60	.916*	-3.001
Descriptive Question	2.00 (0-57)	1.50 (0-10)	1.40 to 2.60	.652*	-.245
Verbal Question	14.00 (0-92)	8.50 (0-94)	9.80 to 18.20^	.840*	-1.425
Reflective Question	3.00 (0-44)	2.00 (0-42)	2.10 to 3.90^	.765*	-1.037
Reflective Statement	13.00 (0-54)	12.00 (0-46)	9.10 to 16.90	.817*	-.283
Statement	43.00 (1-122)	42.50 (1-104)	30.10 to 55.90	.741*	-.649
Descriptive Comment	13.00 (0-72)	10.00 (0-55)	9.10 to 16.90	.807*	-3.365*
Verbal labelling	16.50 (0-114)	15.50 (0-90)	11.55 to 21.45	.890*	-1.039
Indirect Command	58.00 (7-129)	51.00 (16-107)	40.60 to 75.40	.904*	-3.033
Direct Command	31.00 (3-99)	35.00 (4-134)	21.70 to 40.30	.940*	-3.618*
Child Items					
Physical Negative	0.00 (0-16)	0.00 (0-8)	-	.774*	-2.856
Destructive	1.00 (0-29)	1.00 (0-19)	0.70 to 1.30	.565*	-.447
Smart Talk	1.00 (0-17)	0.00 (0-18)	0.70 to 1.30^	.641*	-.041
Cry/Whine/Yell	5.00 (0-57)	8.00 (0-72)	3.50 to 6.50^	.910*	-3.728*
Positive Affect Non-Verbal	14.00 (2-55)	9.00 (0-47)	9.80 to 18.20^	.837*	-3.640*
Positive Affect Verbal	6.50 (0-33)	7.00 (0-25)	4.55 to 8.45	.580*	-1.044
Physical Warmth	0.00 (0-3)	0.00 (0-4)	-	.887*	-2.034

NOTE: Bonferroni correction applied to *p* value = .001

* $p < .001$

^Using the Live median rank scores as the gold standard Video scores exceed the ±30% clinically relevant threshold

Table 3.

Means, standard deviations (*SD*), Pearson's correlations and paired t tests for the four theorized composite variables of the DPICS-R in live and video modes of observation (*N* = 40)

	Live *M* (*SD*)	**Video *M* (*SD*)**	**± 30%**	***r***	***t***
Positive Parenting	57.23 (26.04)	53.83 (23.16)	40.06 to 74.39	.881*	1.742
Negative Parenting	22.30 (18.84)	22.45 (16.57)	15.61 to 28.99	.833*	-.091
Child Positives	29.20 (19.99)	22.68 (14.65)	20.44 to 37.96	.844*	4.016*
Child Negatives	14.95 (16.39)	16.73 (16.54)	10.46 to 19.44	.890*	-1.454

NOTE: Bonferroni correction applied to *p* value = .001

* $p < .001$

Table 4.

Intra-class correlations to assess inter-rater reliability (*N* = 12) of each individual item and composite variable of the DPICS-R using a two-way mixed model with absolute agreement

	Live ICC	***Video ICC***
Parent Items		
Physical Intrusion	.777*	.967*
Physical negative	.198	.897*
Critical Statement	.138	.978*
Negative Command	.474	.963*
Physical Positive	.768*	.967*
Positive Affect	.612	.930*
Unlabeled Praise	.917*	.962*
Labelled Praise	.486	.878*
Acknowledgement	.774*	.917*
Question	.702	.897*
Descriptive Question	.302	.860*
Verbal Question	.931*	.777*
Reflective Question	.679	.811*
Reflective Statement	.913*	.903*
Statement	.927*	.804*
Descriptive Comment	.555	.924*
Verbal labelling	.606	.977*
Indirect Command	.863*	.984*
Direct Command	.944*	.983*
Child Items		
Physical Negative	.723	.563
Destructive	.383	.975*
Smart Talk	.511	.925*
Cry/Whine/Yell	.968*	.996*
Positive Affect Non-Verbal	.682	.982*
Positive Affect Verbal	.401	1.000
Physical Warmth	.706	.949*
Four Theorised Composite Variables		
Positive Parenting	.694*	.958*
Negative Parenting	.764*	.980*
Child Positives	.720	.964*
Child Negatives	.916*	.998*

NOTE: Bonferroni correction applied to *p* value = .001

* $p < .001$

Table 5.
Cronbach alpha's of the four theorized composite variables of the DPICS-R in both live and video modes of observation

Composite Variables	**Live α**	**Video α**
Positive Parenting	.236	.059
Negative Parenting	.601	.436
Child Positives	.293	.127
Child Negatives	.478	.346